View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Purdue E-Pubs

Modeling Human Performance in Restless Bandits with Particle Filters

Sheng Kung M. Yi[1], *Mark Steyvers*[1], *and Michael Lee*[1]

Abstract

Bandit problems provide an interesting and widely-used setting for the study of sequential decision-making. In their most basic form, bandit problems require people to choose repeatedly between a small number of alternatives, each of which has an unknown rate of providing reward. We investigate restless bandit problems, where the distributions of reward rates for the alternatives change over time. This dynamic environment encourages the decision-maker to cycle between states of exploration and exploitation. In one environment we consider, the changes occur at discrete, but hidden, time points. In a second environment, changes occur gradually across time. Decision data were collected from people in each environment. Individuals varied substantially in overall performance and the degree to which they switched between alternatives. We modeled human performance in the restless bandit tasks with two particle filter models: one that can approximate the optimal solution to a discrete restless bandit problem, and another simpler particle filter that is more psychologically plausible. It was found that the simple particle filter was able to account for most of the individual differences.

Keywords

restless bandits, reinforcement learning, sequential decision-making, change detection, non-stationary environments

[1]University of California - Irvine

http://dx.doi.org/10.7771/1932-6246.1060

Modeling Human Performance in Restless Bandits with Particle Filters

Many real-world environments involve temporal changes that require decision-makers to adapt their strategies over time. For example, stock market analysts need to track temporal changes in the market carefully, and sport coaches need to track changes in the performance of a team. In some environments, a decision has to be made between different choices, each of which might be associated with uncertain outcomes that can change over time. For example, drivers have to choose between a number of routes, each associated with some uncertainty about travel times. In addition, traffic changes can lead to changes in the desirability of routes, requiring drivers to adapt their driving strategy continually. In this research, we study how people perform in sequential decision-making situations where each alternative is associated with an uncertain payoff and the underlying environment can change at any time, leading to different payoffs for each alternative.

Bandit problems, as originally described by Robbins (1952), provide a classic task to study sequential decision making. In a standard, stationary bandit environment, people are given a limited number of sequential selections among a fixed set of alternatives, or arms. After each decision, an outcome is generated based on a hidden reward distribution specific to the alternative chosen; the task of the decision-maker is to maximize the total outcomes after all selections have been made. In order to be successful, decision-makers in a bandit environment have to balance their selections between general exploration and exploitation behaviors. Exploration is characterized by the selection of different arms to learn about the hidden outcome distributions for each alternative. Exploitation is characterized by a focus on a single arm, in order to obtain rewards from an option that is believed to be sufficiently good as compared to the other competing options. An expected behavior in a standard bandit problem may start with a period of exploration, followed by exploitation for the remaining choices.

In a standard bandit problem (also called a "game"), the reward rate for each alternative is kept constant over all of the trials. The number of trials in each game may be known, creating a finite horizon problem, or unknown, creating an infinite horizon problem. Optimal solutions can be found for all cases in finite horizon environments by using a dynamic programming approach, where optimal decisions are computed for all potential cases starting from the final trial and solving for each trial toward the first (Kaelbling et al., 1996). As the length of a game increases or the number of alternatives increases, the computation necessary to create a complete decision tree increases exponentially. For infinite horizon problems, certain cases may be solved using Gittins indices (Gittins, 1989). A Gittins index gives each alternative a utility that takes into account an alternative's current estimated value and the information that can be gained from choosing the alternative; the optimal decision is the arm which has the largest index value. However, Gittins indices are only applicable to a limited number of bandit problems, and can be difficult to compute even in those cases (Berry & Fristedt, 1985).

When optimal solutions are available, bandit problems provide an opportunity to examine whether or how people make the best possible decisions. For this reason, many previous empirical studies have been motivated by economic theories, with a focus on deviations from rationality in human decision-making (e.g., Banks, Olson, & Porter, 1997; Meyer & Shi, 1995). More recently, human performance on the bandit problem has been studied within cognitive neuroscience (e.g., Cohen, McClure, &Yu, 2007; Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006) and probabilistic models of human cognition (e.g., Steyvers, Lee, & Wagenmakers, 2009).

The environments in empirical studies have ranged from simple, two-choice bandit problems with either one or two non-deterministic arms (Avineri & Prasher, 2006; Banks et al., 1997; Ben-Elia et al., 2008; Meyer & Shi, 1995) to more complicated environments with more than two probabilistic arms (Steyvers et al., 2009). While there is evidence of significant variation in how people make decisions, people are able to perform significantly better than chance performance, though few are able to match optimal levels of performance.

Most bandit problem research has focused on stationary bandit problem environments, and there has been relatively little focus on the restless bandit problem, especially in empirical work. In the restless bandit problem, the reward rates for alternatives may change over time, rather than remaining stationary through each trial of a game (Whittle, 1988). The introduction of non-stationary outcome distributions adds a large element of complexity in computing optimal decision processes. But it also provides a strong tie to realistic applications, since most sequential decision-making environments found in real life require consideration of changes in the environment. People making decisions in a restless environment are faced with the additional task of change detection (Brown & Steyvers, 2005; Chinnis & Peterson, 1968, 1970; Massey & Wu, 2005), forcing a continuous switch between exploration and exploitation that is not present in the stationary case.

Few empirical studies have looked at human performance in restless bandit tasks. Estes (1984) looked at human performance in repeated two-armed bandit problem games with one known arm and one fluctuating arm. The known arm provided payoffs at a constant probability, while the fluctuating arm provided payoffs with probabilities in a sine-wave pattern over the course of a game. It was found that subjects made choices in a wavelike pattern corresponding to the variation in the alternative reward probabilities. The restless bandit problem has also been studied via brain imaging by Daw et al. (2006), where brain activity has been found to be correlated to obtained rewards and exploratory decision-making. Finally, there are a few animal learning studies that have measured the ability of animals to adapt to changes in reinforcement schedules. These experiments have shown there are substantial individual differences in the ability to track and respond to changes (Gallistel, 2001).

Here, we use a particle filter approach to finding solutions to the restless bandit

problem. Particle filtering is a sequential Monte Carlo method, where a set of particles is updated at each time point to estimate the current state of an environment (see Doucet, de Freitas, & Gordon, 2001). Particles can be thought of as propositions about the environment's current state; as information is gained, particles that describe the observed data well tend to be propagated, while those that do not will tend to be replaced. Over the set of all particles, the propositions form an estimate of the distribution of environment states. These estimates can then be used to make an informed decision on each step of a problem. Particle filters can be used in many situations where Markov Chain Monte Carlo (MCMC) methods become inefficient. For environments where a long history may need to be maintained, the MCMC method will require more computation time with increasing information. In contrast, particle filters, depending on how they are designed, will require less computation time because only a constant set of hypotheses about the current environmental state needs to be maintained.

Particle filters also hold potential in use as descriptors of human performance (Brown & Steyvers, 2009; Daw & Courville, 2007; Sanborn, Griffiths, & Navarro, 2006). By relaxing or changing model parameters, we obtain behaviors that deviate from the optimal strategy in ways that may be useful in describing human performance on restless bandit problems. An early application of particle filters in cognitive science is provided by Sanborn, Griffiths and Navarro (2006), who studied sequential effects in category learning. In their modeling, particles correspond to different mental hypotheses about category structures that the human learner might track. By manipulating the number of particles, their category learning model naturally spanned an interesting range of theoretical possibilities. In particular, when restricted to a single particle, their model reduced to Anderson's (1991) classic "rational model" of category learning, but for a sufficiently large number of particles their model mimicked optimal category learning behavior. In this way, finding the number of particles needed to model people's behavior in sequential category learning tasks provided a natural theoretical mechanism for estimating the complexity of the hypotheses considered by people in learning, and the rationality of their performance.

A second, very recent application of particle filter methods in cognitive science is provided by Brown and Steyvers (2009). These authors applied the particle filter as a descriptor of human performance on an inference and prediction task where the outcome generation distribution changed over time. Individual differences in human performance on the two tasks could be described through shifts in the particle filter's behavior over changes in model parameters. As change detection is a key part of decision-making for the restless bandit, there is potential for application of particle filters to describe individual differences in human behavior for the restless bandit problems as well.

In this paper, we present two different simple restless bandit environments for which particle filter solutions can be employed. We compare these models to the performance of humans in these environments.

Experiment 1

The restless bandit problem used in the first experiment is an extension of sequential stationary infinite-horizon problems. The stationary infinite-horizon bandit problem is one in which, after each decision trial, there is a set probability γ that the game will end. That is, the distribution of individual game lengths follows a Geometric distribution with parameter γ. While games do have an expected length, there is no way of knowing when a game will end. We can obtain the restless bandit environment by considering the scenario where these infinite-horizon games are played consecutively without breaks, such that the indication of the end of each individual game is removed. Without a clear delineation between the change in reward rates, the decision-maker must have a method for noticing these changes in order to maintain good performance. In the standard bandit problem, the shift between exploration behaviors and exploitation behaviors generally occurs only once in a single game. For the restless bandit, we expect a shift back and forth between exploration and exploitation, as periods of stable reward rates are split by changes that must be detected and accommodated in making decisions.

We observe the behavior of human participants in these restless bandit environments, and compared their performance to two different particle filter methods of solutions. One of these solutions is optimal, while the other is sub-optimal but has a more flexible range of possible behaviors.

Participants

Twenty-seven participants drawn from the University of California, Irvine Human Subjects Pool performed the experimental task for course credit. No demographic information was recorded.

Design

Participants played a series of games (blocks) in restless bandit environments. On each trial t of each game, participants were asked to select one of $N = 4$ alternatives. Each selection D_t generated an outcome y_t of either a reward or no reward, based on a Bernoulli draw on the selected alternative's reward rate for that trial $\theta_{D_t,t}$ with $y_t \sim \text{Bernoulli}(\theta_{D_t,t})$.

Reward rates θ on each alternative were drawn from a common generating distribution at the start of each game. On each successive trial, the outcome z_t of a Bernoulli draw with parameter γ determined the alternatives' reward rates: with probability $1 - \gamma$, reward rates on all arms on trial t were maintained to be the same as on trial t-1, and with probability γ, reward rates on all arms were redrawn from the generating distribution. An example of the outcome generation procedure for a three-arm, 20-trial game can be observed in Figure 1. Arm reward rates are visible in the upper plot, changing at randomly distributed time points. Outcomes generated are visible in the lower plot; the only outcomes that are observed are from the arms that are chosen, denoted by the gray boxes.

Figure 1. Illustration of outcomes generated for Experiment 1. Hidden reward rates change at discrete time points and are redrawn from the generating distribution at each change. Outcomes from each arm are generated based on reward rates; only outcomes from selected arms are made visible.

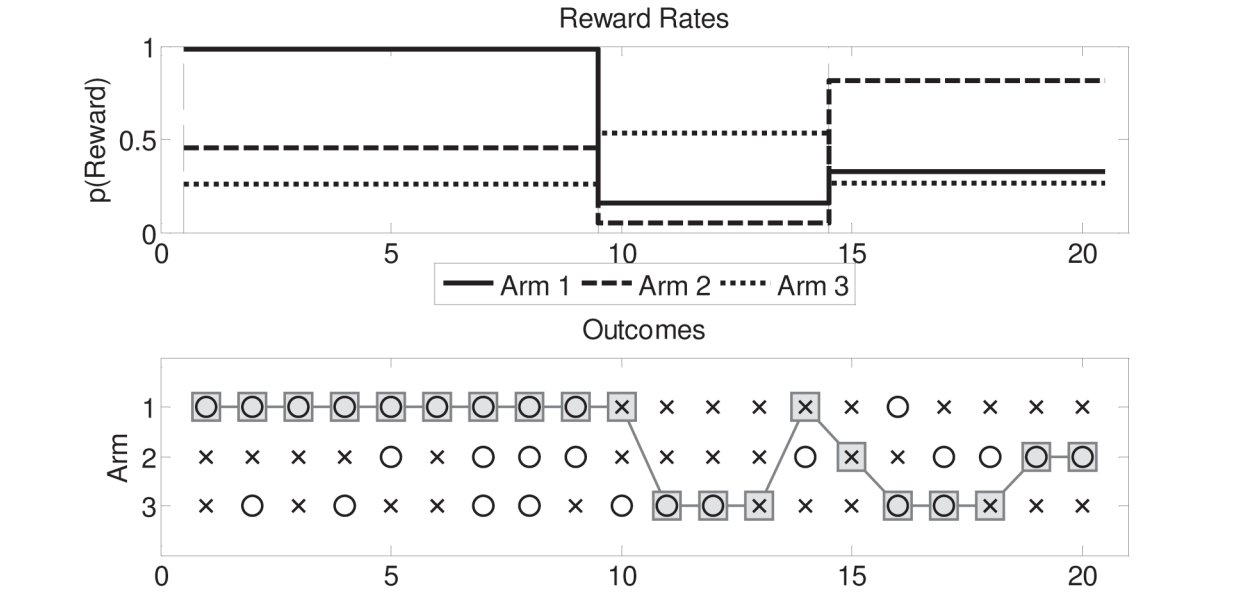
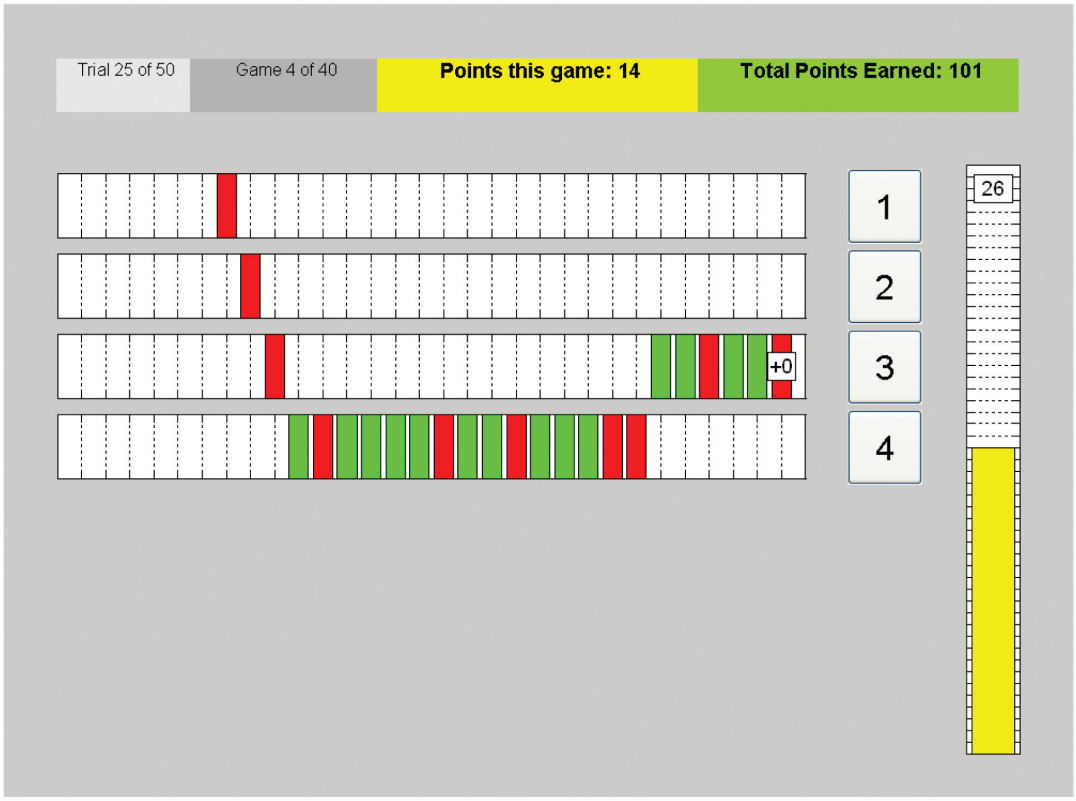


Figure 2. Illustration of the task interface for Experiment 1 and 2.



Each participant played through a total of $G = 42$ games, each with $K = 50$ trials on $N = 4$ alternatives. In all games, the reward rates θ were drawn from a Beta(1,1) distribution (i.e., the uniform distribution) and we set the change rate $\gamma = 0.2$. To facilitate the comparison between individual participants and models, a random seed was set such that all participants went through the same sequence of games, with the same potential rewards for each of the alternatives in each game. The first two blocks performed by each participant were excluded from the final analysis as practice blocks; results reflect the actions performed in only the last forty blocks of the experiment.

Apparatus

The task was performed through a program coded in MATLAB. An example of the experiment interface can be seen in Figure 2. Buttons on the right side of the window represented the alternatives to be chosen, and each selection generated an outcome in the plots to the immediate left of each button. Red bars indicated that no reward was gained, while green bars indicated that a reward was gained. A plot on the far right indicated the number of trials remaining in the current game, while text at the top of the window indicated the current trial, game, and total rewards accumulated. Participant selections and rewards were recorded, as were the hidden reward rates on each trial. Reaction time data were not recorded.

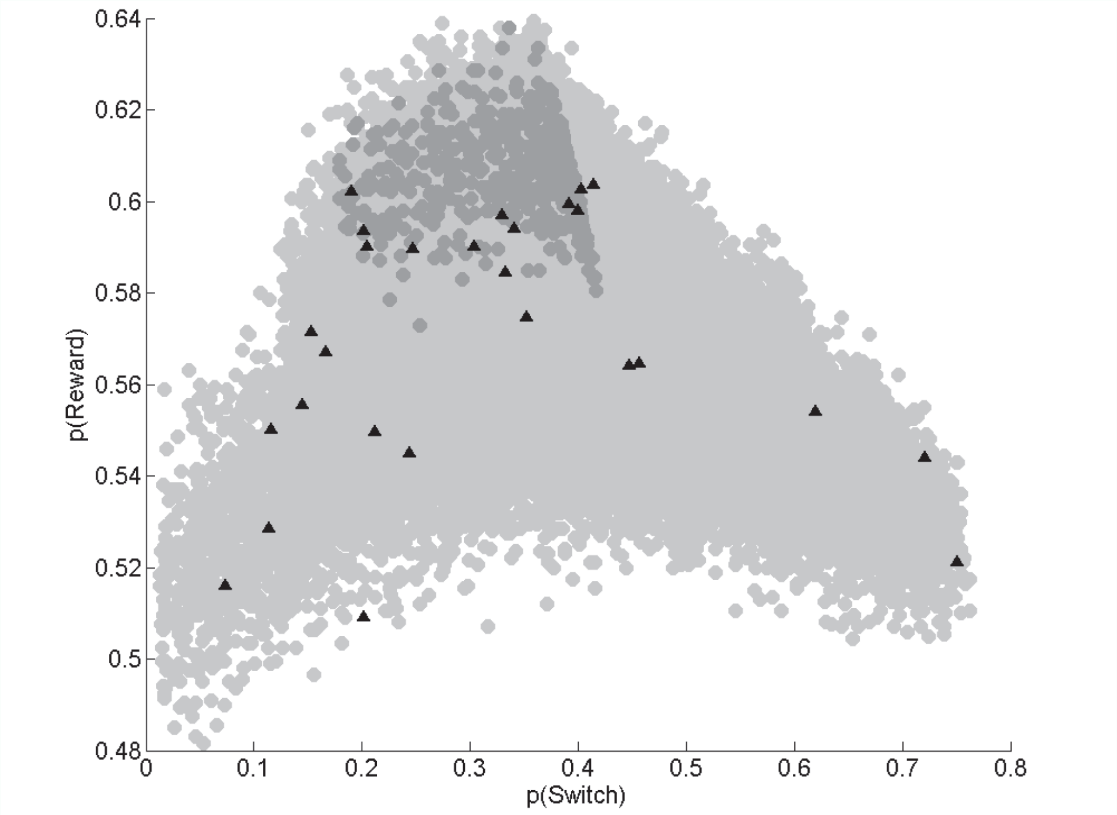
Procedure

Participants were introduced to the task without going into a detailed explanation of the reward-generation process. Participants were simply told that reward rates were randomly generated without describing the precise generating distribution, and that reward rates would change occasionally over the course of each game. Participants were asked to select the alternatives that would maximize their total reward. To maintain focus on the reward maximization objective, a half-second time penalty before the next selection was given if a no reward outcome was generated.

Results

Participant performances over the final forty blocks of the experiment were evaluated with two different measures: the proportion of trials where a reward was obtained, and the proportion of trials where a switch was made. A switch was counted when the alternative chosen in a trial was not the same as that chosen in the previous trial. Figure 3 shows a plot where the performance of each individual participant is marked with a black triangle. It is immediately clear that there is large individual variation in task behavior and performance. A roughly inverted U-shape can be observed, where those with the best performances tend to have a moderate level of switching, while those who switch arms too often or too rarely experience lower performance.

Figure 3. Subject performances in Experiment 1 (black triangles), against the range of the optimal particle filter (dark gray) and reward rate particle filter (light gray).



Optimal Particle Filter

An optimal decision-making procedure can be produced by decomposing the problem into two major components. In the first step, based on decisions and outcomes of previous trials, the probability of a change in reward rates is estimated for each trial. In the second step, these change probability estimates can be translated into a distribution of stable periods where there are no changes in reward rate. An optimal decision can be found for each of these periods. By then aggregating over all possible periods and their posterior probabilities of being the true state, the best decision for the next trial of the game can be calculated.

This general method of solution lends itself naturally to computation via Monte Carlo methods. Particle filtering comes as an especially useful way of looking at this restless bandit problem. Each particle contains a single prediction about which trials are associated with a change in reward rates. These predictions also specify how many trials preceding the current trial are associated with the current reward rates. If we make a prediction of when the next change in reward rate will occur, then we obtain an interval with constant reward rates, turning the problem into a finite bandit problem for which

an optimal solution can be obtained. In this particle filter, we can derive optimal solutions for all inferred stable intervals and pick the alternative that corresponds to the mode of the distribution of all optimal solutions. This approximates, in the limit of the number of particles, the optimal decision for a trial.

Particles that better match the observed data have a higher chance of being propagated to future trials, and so over a sufficiently large pool of particles, an accurate estimate of the changepoint distribution can be obtained. Fewer particles lead to a decrease in the resolution with which the changepoint distribution is approximated and leads to decreases in the model's performance. This may, however, be useful in describing human performance and sub-optimality on the task. In the current implementation of the optimal particle filter, we look not only at the optimal case with sufficiently large number of particles and properly set expected change rate, but also the range of performance over varying numbers of particles and different expected change rates, and compare the optimal model's performance to the range observed in human performance. Varying the number of particles maintained at each trial will have a general effect on reward rate, which may correspond to a general level at which participants are thinking about their decisions. Changing the expected change rate will cause changes in the general behavior of the model that can be related to conservative behavior with few switches between options or liberal behavior with many switches between options found in individuals. Details of the optimal particle filter can be found in Appendix A.

Reward Rate Particle Filter

In the optimal particle filter, particles retain estimates of the trials at which changes in the reward rates may have occurred. Each particle can be used to specify a stable interval where no changes in reward rate occur; each interval has a calculable solution and optimal decision. Over multiple particles, the modal alternative that is chosen will be selected as the optimal decision for a trial. While this can provide an optimal solution, this model might not be psychologically plausible. The decision step relies on an ability to compute the optimal decision for finite bandit problems that is unlikely to be available to human performers.

For these reasons, we also considered an alternative approach to solving the problem, which may be more useful in describing human decision-making behavior. In the reward rate particle filter, particles retain an estimate of the reward rate on each alternative. Over all particles, we obtain an estimate of the current reward rate for each alternative in the form of distributions. The decision step is greatly simplified, taking a greedy approach to selection. For each particle, the alternative with the highest reward rate is taken as the best option; the modal alternative chosen over particles is the model decision for the trial. The changes made to the particle filter model reduce the maximum potential performance, but also increase the range of behaviors observable in a fashion that may

better account for individual differences observed in human participants. As with the optimal particle filter, we have two parameters we manipulate: number of particles, and expected change rate. These parameters affect the model's behaviors in similar fashions as the optimal particle filter. Details of the reward rate particle filter implementation can be found in Appendix B.

Modeling Results

Performance of both the optimal and reward rate particle filters were evaluated in terms of overall reward rate and inter-trial switch probability played over the same forty blocks completed by participants over an array of model parameters. Parameter values ranged from 1 to 200 for number of particles P and from 0 to 1 for expected change rate γ, and performance for each model was evaluated multiple times for each parameter pair. The range of performance under both particle filter models is marked in Figure 3 by the gray shaded areas. The optimal particle filter's range is plotted in dark gray, and the reward rate particle filter is plotted in light gray.

The optimal particle filter's range of performance clearly does not describe the majority of participants well. The model's reliance on optimal decision behavior in the final decision step sets the base reward rate of the model to a level that is comparable to that of the best human performers, even when the model is limited to maintaining only a single particle at each trial. As the number of particles used is increased, performance also increases, and as the internal estimate of the change rate increases, so do the proportion of trials where a switch is made. For higher internal expectations of change rate, model behaviors are similar to a "Win-Stay, Lose-Shift" heuristic strategy: When a reward is obtained on the most recent trial, the same arm is selected with probability p, with a random other arm chosen otherwise; when a reward is not obtained, a random other arm is chosen with probability p, staying on the same arm otherwise. Values of p approaching 1 increase reward gains while reducing the proportion of switch trials.

In comparison to the optimal particle filter, the reward rate particle filter has a much larger range of overall behaviors over the range of parameters. Increases in the number of particles dramatically raise the reward rate, though the benefits begin to asymptote after approximately 100 particles. Performance of the reward rate particle filter is comparable to that of the optimal at the upper limit; the "greedy" strategy of selecting the arm with the highest expected reward rate does not differ significantly in overall reward rate from the strategy employed by the optimal model.

The fact that both the optimal and reward rate particle filters seem to asymptote at such low numbers of particles is interesting. Despite the risk of degeneracy, performance does not suffer from a relatively coarsely-estimated distribution of change points (in the optimal) or current reward rates (in the non-optimal). It requires relatively little effort to obtain a strategy that performs well above random chance; a pure "Win-Stay, Lose-Shift"

heuristic strategy (where $p = 1$) performs nearly as well as the particle filter models at peak parameter settings, in terms of overall reward rate. In addition to a "greedy" strategy creating little difference in terms of overall performance, the information that needs to be kept in order to make an informed decision does not have to be particularly large. The effect of increasing the number of particles lies mostly in the variability in overall reward rates, where randomness and chance has a larger effect on the number of rewards obtained when fewer particles are maintained between trials.

Despite the fact that simple strategies can create good performance, the fact remains that our empirical results show human decision-makers follow a wide range of behaviors, achieving varying degrees of success. Heuristic strategies and optimal decision-making models are too narrow to account for these individual differences well. The sub-optimal particle filter model, however, perhaps has the flexibility to describe individual performances with intuitively interpretable parameters, while also maintaining the ability to perform at near optimal levels with certain parameter choices. Still, it is difficult to associate best-fitting parameter values to individuals due to the amount of variability present in behavior and performance of the model at each parameter pair. Most participants' performances fall below the asymptotes, and so are better described with fewer particles; with fewer maintained particles there is more variability in performance for the model.

Experiment 2

In Experiment 1, we found that the performance of the reward rate particle filter was, over the range of parameter values, more adept at describing a larger variety of potential behaviors, including those of human subjects, than that of an optimally designed particle filter. In addition, there was relatively little loss in maximum expected rewards when using the reward rate particle filter as compared to the optimal. The reward rate particle filter also carries the advantage that the decision step is much simpler and the information contained in individual particles does not require a memory of previous trials. These properties give the reward rate particle filter the potential for applicability in a wider array of environments than the one for which it was originally designed.

For Experiment 2, we investigate a different restless bandit environment where changes occur continuously over time, rather than at randomly spaced intervals. On each trial, the reward probability on each arm is re-sampled from a distribution centered about the reward probability from the previous trial. Thus, while there is little change in reward rates from trial to trial, over multiple trials, the reward rates may change drastically. In this environment, the best strategy to employ at each time point is to choose the alternative with the largest estimated reward rate.

Once again, we observe the behavior of human decision-makers in this environment and compare them to particle filter models. Here, we employ two types of reward rate

particle filters: one that has a propagation mechanism that matches the environment and a second identical to the one used in Experiment 1 that does not match the environment. Our goal is to observe how much loss the use of the inappropriate reward rate particle filter incurs, and how it compares to the optimal reward rate particle filter.

Participants

Thirty-six participants were drawn from the University of California, Irvine Human Subjects Pool to perform the experimental task, for course credit. No demographic information was recorded.

Design

All aspects of the experimental design were maintained from Experiment 1 except for the reward rate generation mechanism, such that changes in reward rates occurred continuously over trials, rather than at discrete time points. On the first trial of each game, reward rates were drawn from a common generating distribution Beta(1,1), as in Experiment 1. On each subsequent trial, a new reward rate was drawn on each arm dependent on the previous reward rate, $\theta_{i,t} \sim \text{Beta}(1+c(\theta_{i,t-1}),1+c(1-\theta_{i,t-1}))$, where c is a parameter that controls the variability in reward rate between trials. The mean of the distribution, $(1 + c(\theta i,t-1))/(2+c)$, is slightly biased toward the neutral value of 0.5. Increasing parameter c decreases the variance in the generating distribution and shifts the mean toward that of the previous trial's reward rate. An example of a three-arm, 20-trial game can be observed in Figure 4. As with Figure 2, arm rates are observable in the upper panel, while outcomes can be seen in the lower panel. As with Experiment 1, each participant played through the same sequence of $G = 42$ games, each with $K = 50$ trials and $N = 4$ alternatives with $c = 10$. The first two blocks performed by each participant were excluded from the final analysis as practice blocks; results reflect the actions performed in only the last forty blocks of the experiment.

Apparatus

Aside from the reward rate generation method, no changes were made to the experimental program.

Procedure

The procedure was nearly identical to that of Experiment 1. Only a slight change to experimental instructions was made to reflect the changes in reward rate distribution; changes were stated to occur gradually throughout each game, rather than occasionally at random times.

Results

Participant performance was again evaluated on the final forty blocks played, in terms of the proportion of trials where a reward was obtained, and the proportion of trials where a switch was made. Figure 5 shows participant performance on these measures, marked

Figure 4. Illustration of outcomes generated for Experiment 2. Hidden reward rates change after each trial, based on the reward rate on the previous trial. Outcomes from each arm are generated based on reward rates; only outcomes from selected arms are made visible.

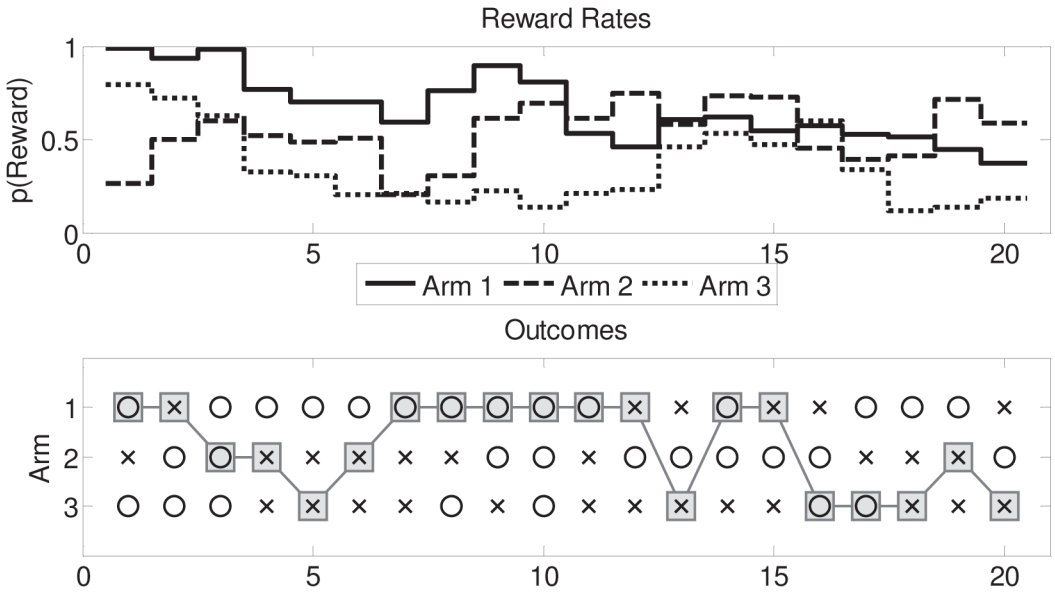
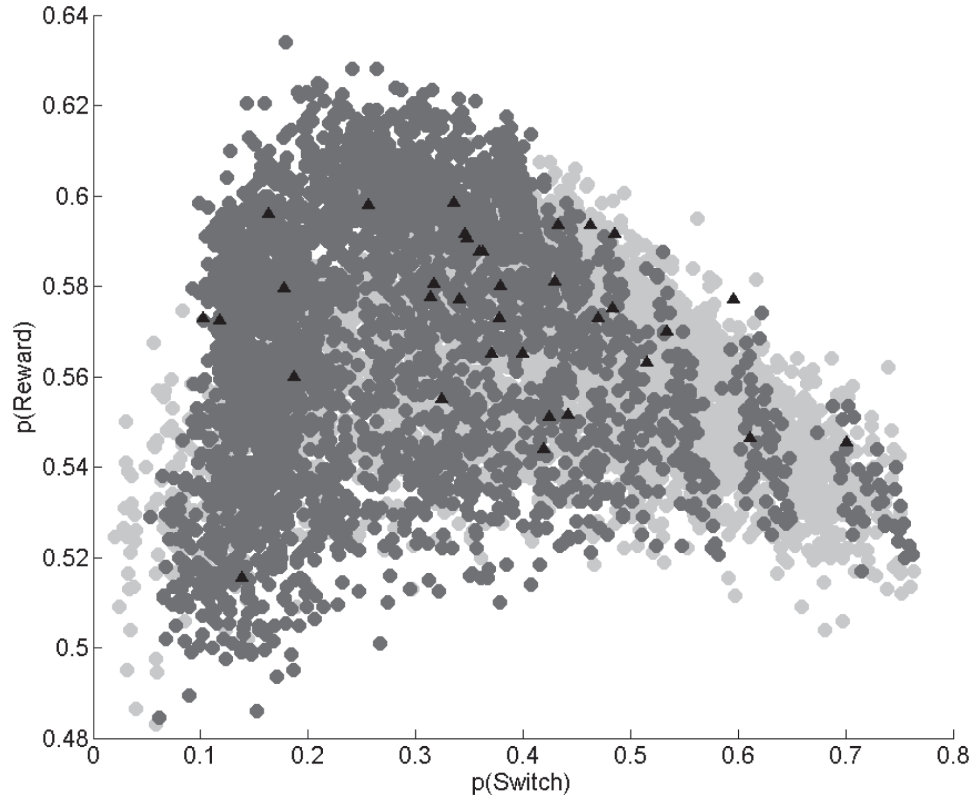


Figure 5. Subject performances in Experiment 2 (black triangles) against the range of the continual-change reward rate particle filter (dark gray) and discrete-change reward rate particle filter (light gray).



with black triangles. As with Experiment 1, there are large individual differences in participant behavior. The inverse U-pattern observed in the previous experiment is again visible, although it is not quite as distinct. A large group of participants with switch proportions between approximately 0.3 and 0.45 have overall behaviors that are in line with the "Win-Stay, Lose-Shift" strategy.

Continual-change Reward Rate Particle Filter

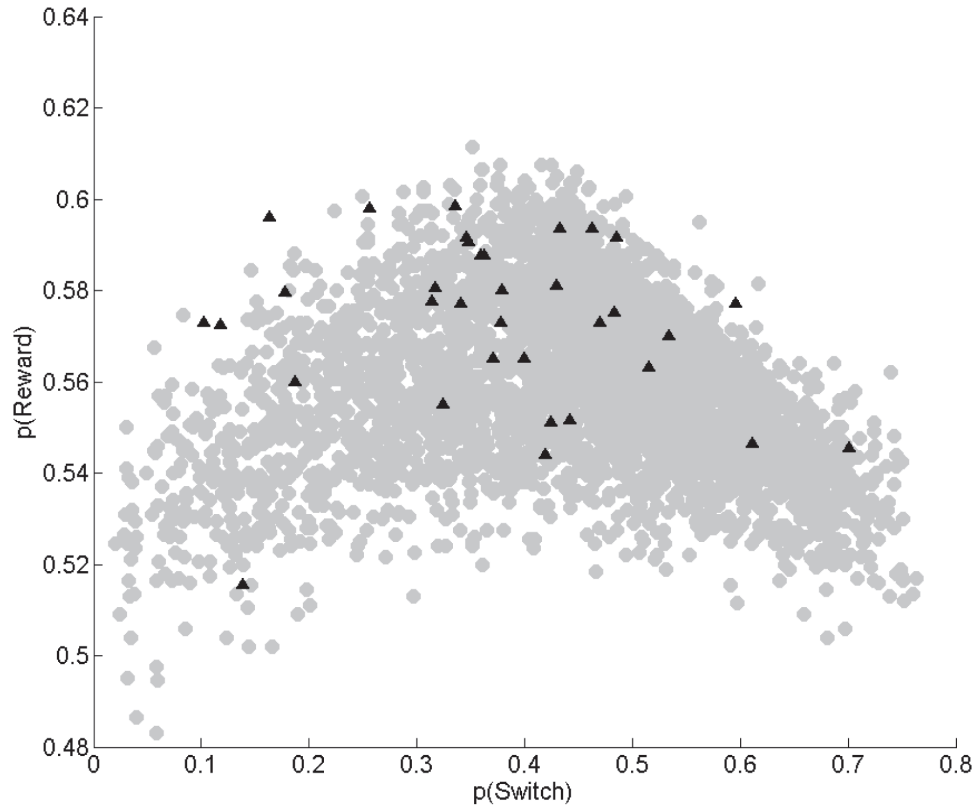
With the change in the reward-generation environment, changes in decision-making strategy must also be made to maintain good performance. A particle filter in which particles carry information regarding reward rates should have each particle adjust its reward rate estimates in the same way as the environment. Sampled particles' estimated reward rates will first be re-sampled from distributions estimating the level of expected drift in trial-wise reward rate before being potentially propagated to the next trial. The decision step of the adjusted particle filter remains the same as before, taking the alternative with the highest estimated reward rate. This particle filter has two parameters that can be varied. The first, the number of particles, serves the same purpose as in the models for Experiment 1 and generally affects the overall performance of the model. The second parameter, representing the expected variability in reward rates between trials, serves a similar function to the expected change rate in the Experiment 1 models and has a general effect on how often the model will tend to switch between arms. Details of the second reward rate particle filter can be found in Appendix C.

Modeling Results

Performance measures were obtained for the continual-change and discrete-change reward rate particle filter models across model parameters in terms of reward rate and switch rate. The range of parameter values tested for the discrete-change particle filter were identical to those in Experiment 1; parameter values for the continual-change particle filter were evaluated over an array covering the range from 1 to 200 for number of particles P and from 0 to 200 for the estimate of variability between trials c. Figures 5 and 6 show the range of performance under both particle filter models, marked by the gray shaded areas. Figure 5 includes both the discrete-change (dark gray) and continual-change (light gray) particle filters, while Figure 6 includes the discrete-change particle filter alone.

The continual-change reward rate particle filter model shows a wide range of overall behaviors across its parameters. Performance changes in an expected pattern: increases in the number of particles result in an increase in the overall reward rate; decreases in the expected variability of the environment reward rates results in a decrease in the amount of switching between alternatives. The discrete-change particle filter shows the same pattern as in Experiment 1, though flattened both due to the difficulty of the continuous-change environment as well as the mismatch between propagation method and environment generation. Comparing the performance of the discrete-change particle filter to that

Figure 6. Subject performances in Experiment 2 (black triangles) against the range of the discrete-change reward rate particle filter (light gray).



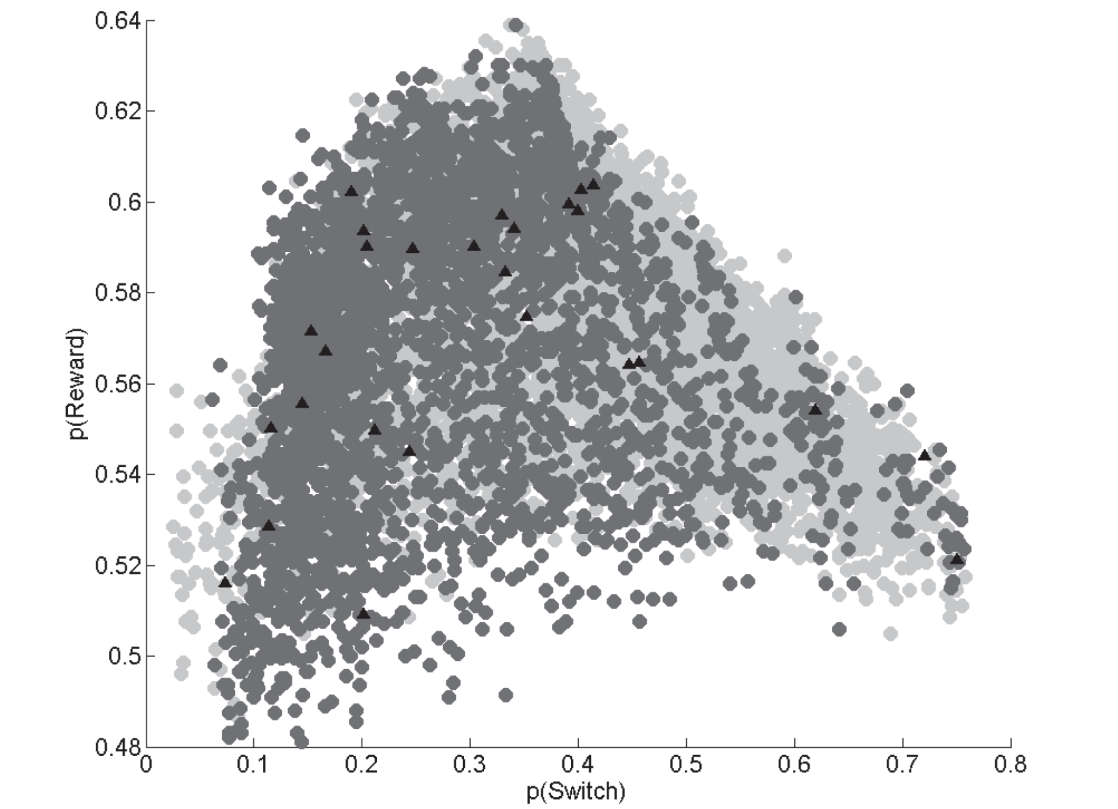
of the continual-change particle filter, we find that at the highest particle counts, the continual-change particle filter does markedly better than the discrete-change model. As a result, there are participants with lower switch rates and high reward rates who are not well-described by the discrete-change particle filter, but fall under the range covered by the continual-change model.

We also compare the continual-change model to the discrete-change model in the environment of Experiment 1. Figure 7 plots the discrete-change (dark gray) and continual-change (light gray) particle filter together in the discrete change-point environment across the range of model parameters. While in Experiment 2, the discrete-change particle filter lost performance due to environment change, there is remarkably little difference between the two models' overall performance range in the Experiment 1 environment.

Discussion

Bandit problems have been utilized extensively in sequential decision-making research, but relatively little empirical research has been done with restless environments, where reward rates may change over time. Here, we have observed the performance of people in two restless bandit environments with different reward rate change dynamics. In the

Figure 7. Subject performances in Experiment 1 (black triangles) against the range of the discrete-change reward rate particle filter (light gray) and continual-change reward rate particle filter (dark gray).



first, reward rates changed at discrete but random time points for all arms simultaneously; in the second, reward rates changed continuously such that there were small short-term changes and large long-term changes. We found that people were able to perform the task in both environments but we also found substantial individual differences in behavior. Generally, participants who performed best were those who switched options at an appropriate rate, while those who switched too much or too little performed comparatively worse.[1] This range of behaviors was well described by particle filter models. Each particle filter that we considered had easily-interpretable parameters. One parameter is the number of particles that modulates the overall performance. The number of particles can be likened to the amount of cognitive resources applied to solving a problem. As the number of particles increases, performance increases with diminishing returns for large numbers of particles. The other parameter related to the perceived variability in the environment and modulated the amount of switching between alternatives. For those particle filter models whose particles contained hypotheses about the reward rates on each option, we found that varying these parameters explained most of the observed individual differences. In

contrast, the optimal model for the first task had a far too narrow range of behaviors to adequately describe individual differences.

A natural application of the particle filter approach would be to assess the best-fitting parameters for individual participants. This would allow a natural explanation of their behavior in terms of interpretable parameters. However, the probabilistic nature of the particle filter model creates a lot of variability in overall performance, making it difficult to obtain a precise estimate of the average behavior for a specific parameter pair. Thus, it is hard to give a precise characterization of a person's performance in terms of a best-fitting particle filter model. While there is a clear relationship between parameter values and behavior, and parameters do have easily interpretable meanings, there is quite a bit of variability in what can be expected from the model's performance. In addition, it is difficult to distinguish between variations of the particle filter model in most cases; there was a lot of overlap in the range of behaviors described by the discrete-change reward rate particle filter and its continuous-change variant, in both experimental environments.

Still, we can note some interesting patterns between the models' performances and that of the human observers. A large number of participants tended to switch at a rate slower than that of the optimal performer. This seems to match some results found in Meyer & Shi (1995) where participants tended to undersample in most cases. It is reasonable to guess that participants will tend to switch less because of a belief of security in staying on an option they know more about over taking risks by exploring a lesser-known option, potentially 'losing' a good, steady reward rate. This translates into performances that match up to model behaviors that assume slower environmental change rates.

Overall, particle filters provide a promising modeling framework not only to approximate the optimal solution in restless bandit environments, but also to model individual differences.

Appendix A

Details of Optimal Particle Filter for Experiment 1

The optimal particle filter solution to the given restless bandit environment follows two major phases. In the first phase, we propagate particles from the previous trial to the current trial, after taking into account the most recently observed outcome. The current implementation of the particle filter uses a direct simulation method, using the following steps:

For $t = 1$, initialize particles $k = 1,...,P$, $\mathbf{z}_1^k \sim Bernoulli(\gamma)$ for each arm, where γ is the expected probability of a change in reward rates.

For $t = 2,\ldots,K$,

Initialize counter $p = 0$.

While $p < P$,

Take sample $k \sim U[1,...,P]$.

Generate proposal $\hat{\mathbf{z}} = \{\mathbf{z}_{t-1}^{k}, z_t \sim Bernoulli(\gamma)\}$ on each arm.

Sample $u \sim U[0,1]$.

If $P(y_t \mid \hat{\mathbf{z}}) > u$, $p = p + 1$; $\mathbf{z}_t^p = \hat{\mathbf{z}}$.

In the second phase, each particle is used to describe an interval where no changes in reward rates occur that includes the current trial. The trial of the most recent change point forms the start of the interval, while a draw from a *Geometric*(γ) specifies how many trials remain before the end of the period. These intervals describe finite-horizon bandit problems, whose solutions can be obtained by dynamic programming (Kaebling et al., 1996). Each interval has an optimal selection for the current trial; the mode selection over all particles is the optimal choice for the model. Two parameters can be manipulated, the number of particles *P* that are maintained on each trial, and the expected change rate γ.

Appendix B

Details of the Discrete-Change Reward Rate Particle Filter

As with the optimal particle filter, each trial's decision is based on a two-step solution of propagating particles, then selecting an alternative based on the mode response over particles. The first step is very similar to that of the optimal particle filter, propagating particles through a direct simulation method, except that particles carry reward rate information at the most recent trial, rather than change point information over the game played so far:

For $t = 1$, initialize particles $k = 1,\ldots,P, \boldsymbol{\theta}_1^k \sim Beta(1,1)$ for each arm

For $t = 2,\ldots,K$,

Initialize counter $p = 0$.

While $p < P$,

Take sample $k \sim U[1,...,P]$.

Generate proposal $\hat{\boldsymbol{\theta}}$ on each arm:

If *Bernoulli*(γ) = 1, $\hat{\boldsymbol{\theta}} \sim Beta(1,1)$ for each arm,

Otherwise $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_{t-1}^k$.

Sample $u \sim U[0,1]$.

If $P(y_t \mid \hat{\boldsymbol{\theta}}) > u$, $p = p + 1$; $\boldsymbol{\theta}_t^p = \hat{\boldsymbol{\theta}}$.

The decision step is considerably simpler than in the optimal model. Each particle gives a predicted reward rate on each arm; the best decision implied by each particle is the alternative with the largest reward rate. The mode alternative over all particles is the choice made by the model for the next presented trial. As with the optimal particle filter, we can manipulate two parameters, the number of particles *P* that are maintained on each trial, and the expected change rate γ.

Appendix C

Details of Continual-Change Rate Particle Filter

The continual-change rate particle filter operates similarly to the discrete change particle filter, with an identical decision method. What differs is the particle propagation method—after a particle is sampled, the proposal particle reward rates are sampled from a distribution centered around the reward rates of the sampled particle:

For $t = 1$, initialize particles $k = 1,\ldots,P$, $\boldsymbol{\theta}_1^k \sim Beta(1,1)$

For $t = 2,\ldots,K$,

Initialize counter $p = 0$.

While $p < P$,

Take sample $k \sim U[1,...,P]$.

Generate proposal $\hat{\boldsymbol{\theta}}$ on each arm:

$$\hat{\theta}_i \sim Beta(1 + c(\theta_{i,t-1}^k), 1 + c(1 - \theta_{i,t-1}^k)) \text{ for } i = 1,\ldots,N.$$

Sample $u \sim U[0,1]$.

If $P(y_t \mid \hat{\boldsymbol{\theta}}) > u$, $p = p + 1$; $\boldsymbol{\theta}_t^p = \hat{\boldsymbol{\theta}}$.

As with the discrete-change reward rate particle filter, the alternative chosen at the decision step is the mode decision over all particles, where each particle implies the best decision to be on the arm with the largest reward rate. There are two parameters that we can manipulate, the number of particles P that are maintained on each trial, and the estimate of variability in reward rates between trials c. It should be noted that increasing parameter c serves to decrease the variance in reward rate when a particle is propagated.

Endnote

1. One may wonder if probability matching can explain the observed switching behavior. When faced with sequential decision-making tasks, there seems to be a tendency for agents to select between options with a distribution that reflects the relative reward rate of each option, whether it is the correct policy (Baum, 1975; Gallistel et al., 2001) or not (Tversky & Edwards, 1966; West & Stanovich, 2003). However, it is unclear how probability matching applies in the restless bandit problem because performers generally did not switch frequently enough to suggest probability matching as a strategy being used.

References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98*(3), 409-429.

Avineri, E. & Prashker, J. N. (2006) The impact of travel time information on travelers' learning under uncertainty. *Transportation*, *33*(4), 393-408.

Banks, J., Olson, M., & Porter, D. (1997) An experimental analysis of the bandit problem. *Economic Theory*, *10*, 55–77.

Baum, W. M. (1975) Time Allocation in Human Vigilance. *Journal of the Experimental Analysis of Behavior*, *23*, 45-53.

Ben-Elia, E., Erev, I., Shiftan, Y. (2007) The combined effect of information and experience on drivers' route-choice behavior. *Transportation*, *35*(2), 165-177.

Berry, D. A., & Fristedt, B. (1985) *Bandit Problems: Sequential Allocation of Experiments*. New York, Chapman and Hall.

Biele, G., Erev, I., Ert, E. (in press). Learning, risk attitude and hot stoves in restless bandit problems. *Journal of Mathematical Psychology*.

Brown, S., Steyvers, M. (2009). Detecting and predicting changes. *Cognitive Psychology*, *58*, 49-67.

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? exploration versus exploitation. *Philosophical Transactions of the Royal Society B: Biological Sciences, 362*, 933–942.

Chinnis, J., & Peterson, C. (1968) Inference About a Nonstationary Process. *Journal of Experimental Psychology*, 77, 620-625.

Chinnis, J., & Peterson, C. (1970) Nonstationary Processes and Conservative Inference. *Journal of Experimental Psychology*, 84, 248-251.

Daw, N. D., & Courville, A. C. (2007). The rat as particle filter. Advances in Neural Information Processing, 20, 369–376.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006) Cortical substrates for exploratory decisions in humans. *Nature*, *441*, 876–879.

Denrell, J., & March, J. G. (2001) Adaptation as Information Restriction: The Hot Stove Effect. *Organization Science*, *12*(5), 523-538.

Doucet, A., de Freitas, N., & Gordon, N. (2001). Sequential Monte Carlo methods in practice. New York: Springer.

Estes, W. K. (1984) Global and Local Control of Choice Behavior by Cyclically Varying Outcome Probabilities. Journal of Experimental Psychology: Learning, Memory, and Cognition, 10(2), 258-270.

Gallistel, C. R., Mark, T., King, A., & Latham, P. E. (2001). The rat approximates an ideal detector of changes in rates of reward: Implications for the law of effect. Journal of Experimental Psychology: Animal Behavior Processes, 27, 354–372.

Gittins, J. C. (1989). *Multi-armed bandit allocation indices*. New York: Wiley.

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996) Reinforcement Learning: A Survey. *Journal of Artificial Intellignce Research*, *4*, 237-285.

Massey, C., Wu, G. (2005) Detecting Regime Shifts: The Causes of Under- and Overreaction. *Management Science*, 51(6), 932-947.

Meyer, R. J., Shi, Y. (1995) Sequential Choice Under Ambiguity: Intuitive Solutions to the Armed Bandit Problem. *Management Science*, *41*(5), 817-834.

Robbins, H. (1952). Some aspects of the sequential design of experiments. Bulletin of the American Mathematical Society, *55*, 527–535.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2006). A more rational model of categorization. *Proceedings of the 28th Annual Conference of the Cognitive Science Society.*

Steyvers, M., Lee, M.D., & Wagenmakers, E.J. (2009) A Bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*, *53*, 168-179.

Tversky, A., Edwards, W. (1966) Information Versus Reward in Binary Choices. *Journal of Experimental Psychology*, *71*(5), 680-683.

West, R. F., Stanovich, K. E. (2003) Is probability matching smart? Associations between probabilistic choices and cognitive ability. *Memory & Cognition*, *31*(2), 243-251.

Whittle, P. (1988). Restless Bandits: Activity Allocation in a Changing World. *Journal of Applied Probability, 25*A, 287–298.

Acknowledgments

This work was funded by award FA9550-07-1-0082 from the Air Force Office of Scientific Research.

Paper submitted on March 26, 2009.

The final version accepted on August 14, 2009.